Special issue on Variable Selection and Robust Procedures

Modern technology makes it much easier to record and store a lot of information about subjects, processes or phenomena under study. As a consequence, in empirical research one is often confronted with a wealth of variables. On the other hand, since data storage has become so cheap, in some applications information on a huge number of subjects is readily available. In both cases, the underlying data generating mechanism is most likely very complex. A key data analysis target is to understand the most relevant relations among the variables. While the wealth of available data may seem as a blessing, it may also present difficult challenges. For example, more complex statistical models – closely resembling the underlying data generating process – can be fit when larger data sets are available. However, this usually implies that an optimal model needs to be selected from a flexible and thus large class of models. Furthermore, when more variables are recorded, it becomes more likely that not all of these variables are measured with high accuracy. This may result in data of uneven quality that contains gross errors and other anomalies. Such deviations need to be taken into account in an appropriate analysis of the data. Hence, the methods and techniques that are used to analyze large data sets need to be robust against data anomalies. Since usually a large number of candidate models needs to be fitted, computational efficiency is a major concern. Indeed, for complex methods it may become infeasible to evaluate all candidate models due to their high computational cost.

This special issue focuses on the following two challenges: (a) model selection strategies, methodology and applications, and (b) computationally efficient, robust procedures to analyze complex data sets. The journal 'Computational Statistics and Data Analysis' has paid much attention to these important topics in recent years. In 2007 a special issue on 'Machine Learning and Robust Data Mining' focused on machine learning and robust data mining techniques to analyze huge complex data sets (Croux et al., 2007). At the same time a special issue on "Statistical Learning Methods Including Dimensionality Reduction" focused on model selection (Bock and Vichi, 2007). Also more recently, many articles focusing on robustness have appeared in this journal (see e.g. Hubert et al., 2009; Frahm and Jaekel, 2010; Harrington and Salibián-Barrera, 2010; Lanius and Gather, 2010; Schyns et

al., 2010; Todorov and Filzmoser, 2010; Wiens and Wu, 2010). The problem of variable selection has also been considered in many contributions (Daye and Jeng, 2009; Lu, 2009; Maugis et al., 2009; Wang, 2009; Ogura, 2010). A few articles focused on the combined problem of variable selection and robustness (Lutz, et al., 2008; Salibian-Barrera and Van Aelst, 2008).

Modeling data distributions is at the heart of all statistical analysis. Not surprisingly, several papers in this special issue focus on robustness and selection problems that arise when modeling a data distribution. Alvarez-Esteban et al. (2010) consider robust tests for normality while Meintanis and Tsionas (2010) propose a goodness-of-fit test to investigate the appropriateness of the generalized normal-Laplace class of distributions to model the data. Rozenholc et al. (2010) focus on the construction of a histogram to visualize the data distribution. Three papers use mixture distributions to model the data. Basso et al. (2010) consider scale mixtures of skew-normal distributions to model irregular data and Abanto-Valle et al. (2010) apply scale mixtures of normal distributions in the stochastic volatility model. Rufo et al. (2010) focus on the calculation of Bayes factors to handle the model selection problem in mixture models.

Several papers in this issue focus on the difficult computational problems that arise when computing robust regression estimators. Indeed, the exact computation of these estimates are often NP-hard problems and so approximate algorithms are needed in practice. Flores (2010) uses approximate reweighted least squares solutions and clustering techniques to efficiently calculate robust regression estimates while Nunkesser and Morell (2010) use evolutionary search heuristics to find robust regression solutions efficiently. A third paper by Nguyen and Welsch (2010) solves semi-definite programming problems to identify outliers and find robust regression estimates.

Robust regression is also the topic of three other papers in this issue. Maronna and Yohai (2010) show that the standard choices of the tuning constants for MM-estimators are not satisfactory for high-dimensional data and propose better choices for these constants in such cases. Adrover and Salibian-Barrera (2010) propose a general method to construct globally robust confidence intervals for the parameters in a simple linear regression model. Jurečková et al. (2010) show how distribution free rank and regression rank score tests can be constructed in the linear model in the presence of measurement error.

A new procedure to select variables in linear regression is investigated by Min

et al. (2010). Methods for variable selection in regression in the presence of outliers are proposed by Menjoge and Welsch (2010) and Riani and Atkinson (2010). Both proposals are based on the forward search procedure. The first paper incorporates outlier identification into the model selection procedure by using dummy variables. The second uses Mallows' C_p for model selection. Evaluating the prediction performance of selected regression models is considered by Borra and Di Ciaccio (2010) who present a detailed comparison of several techniques. Khan et al. (2010) consider the problem of measuring prediction performance of regression models in the presence of outliers.

More general linear models are often used to model and analyze data. For example, generalized linear models are needed when the response is not continuous or mixed models are used to handle repeated measures. In these more complex models, the issues of robustness and model selection are as relevant as in linear regression. Powers et al. (2010) propose a Bayesian approach for variable selection in Poisson regression with underreported responses. Bayesian model selection is considered by Nott and Leng (2010) for generalized linear models and by Overstall and Forster (2010) for generalized linear mixed models. Ouysse and Kohn (2010) use Bayesian variable selection and model averaging for the arbitrage pricing theory model. Model selection based on cross-validation for marginal structural models in the context of causal inference is studied by Haight et al. (2010) while procedures for covariate and model selection when modeling longitudinal data are proposed by Wang and Hin (2010). Boente and Rodriguez (2010) propose a robust estimation procedure for generalized partially linear models and provide a robust Wald type test for the regression parameter in this model.

Multivariate statistical models play a central role in analysis multivariate data. It is well-known that classical estimators based on the multivariate normality assumption are highly sensitive to contamination. Here also the problem of model selection naturally arises. For example, selecting the number of components in a principal component analysis or a factor analysis is a key issue in multivariate data analysis. Detecting influential observations in principal component analysis (PCA) is considered by Boente et al. (2010) and Debruyne et al. (2010). The first paper focuses on PCA and common PCA while the latter considers kernel PCA. A new robust exploratory factor analysis method is proposed by Unkel and Trendafilov (2010) while Chen et al. (2010) use the concept of generalized degrees of freedom to select the number of factors in this context. For confirmatory factor analysis ro-

bust methods based on S-estimators are considered in Dupuis Lozeron and Victoria-Feser (2010). Lykou and Whittaker (2010) use a lasso type approach to obtain a sparse solution for high-dimensional canonical correlation analysis. García-Escudero et al. (2010) consider a robust approach based on trimming to find clusters of observations concentrated around different linear patterns. The dependence structure in multivariate data is investigated through concentration graph models by Gottard and Pacillo (2010) who use the minimum covariance determinant estimator to obtain a robust solution. Emura et al. (2010) use Archimedean copula models to model dependence in right censored data.

Another important issue in statistical data analysis is the presence of missing data. Two papers in this special issue focus on handling missing data. Hron et al. (2010) investigate the use of imputation techniques for missing data when analyzing compositional data. At the same time, robust regression methods are used to handle possible outliers. Schomaker et al. (2010) investigate the performance of model averaging as an alternative to model selection in the presence of missing data.

Smoothing is a popular and generally successful technique to model curvilinear data as encountered for example with spectra or time series. Since smoothing methods are data adaptive it is not surprising that they can be highly influenced by outliers. Croux et al. (2010) proposes a robust smoothing method for multivariate time series based on robust estimation of covariance matrices. Lee and Cox (2010) consider robust smoothing techniques for analyzing spectroscopy data and focus on selecting the smoothing parameter through a computationally efficient and robust cross-validation procedure.

We are confident that the high quality papers in this special issue clearly show the importance of robustness and model selection in contemporary and future statistics and data analysis. The importance of these problems is also illustrated by the large number of submission that we received for this special issue. We greatly acknowledge the help of the CSDA co-editors and several associate editors to handle all these papers, namely S.P. Azen, A. Colubi, P. Duchesne, C. Gatu, M.A. Gil, K. Hornik, M. Hubert, E.J. Kontoghiorghes, K. Kurihara, J.C. Lee, J.J. Lee, M. Mittlböck, I. Moustaki, R. Paap, D.S.G. Pollock, T. Proietti, M.W. Trosset, P. Vieu, H. Wang, J. Whittaker, R. Wilcox, P. Winker, M. Xie, and P.L.H. Yu. Finally, we would like to thank all anonymous referees for their honest opinions, valuable comments and

helpful suggestions.

References

Abanto-Valle, C.A., Bandyopadhyay, D., Lachos, V.H. and Enriquez., I., 2010. Robust Bayesian analysis of heavy-tailed stochastic volatility models using scale mixtures of normal distributions. Comp. Statist. Data Anal. (this issue), doi:10.1016/j.csda.2009.06.011.

Adrover, J. and Salibian-Barrera, M., 2010. Globally robust confidence intervals for simple linear regression. Comp. Statist. Data Anal. (this issue), doi:10.1016/j.csda.2009.05.005.

Alvarez-Esteban, P.C., del Barrio, E., Cuesta-Albertos, J.A. and Matrán, C., 2010. Assessing when a sample is mostly normal. Comp. Statist. Data Anal. (this issue), doi:10.1016/j.csda.2009.12.004.

Basso, R.M., Lachos, V.H., Cabral, C.R.B. and Ghosh, P., 2010. Robust mixture modeling based on scale mixtures of skew-normal distributions. Comp. Statist. Data Anal. (this issue), doi:10.1016/j.csda.2009.09.031

Bock, H.-H. and Vichi, M., 2007. Statistical learning methods including dimensionality reduction. Comp. Statist. Data Anal., 52, 370-373.

Boente, G., Pires, A.M. and Rodrigues, I.M., 2010. Detecting influential observations in principal components and common principal components. Comp. Statist. Data Anal. (this issue), doi:10.1016/j.csda.2010.01.001.

Boente, G. and Rodriguez, D., 2010. Robust inference in generalized partially linear models. Comp. Statist. Data Anal. (this issue), doi:**********.

Borra, S. and Di Ciaccio, A., 2010. Measuring the prediction error. A comparison of cross-validation, bootstrap and covariance penalty methods. Comp. Statist. Data Anal. (this issue), doi:10.1016/j.csda.2010.03.004.

Chen, Y.-P., Huang, H.-C. and Tu, I.-P., 2010. A new approach for selecting the number of factors. Comp. Statist. Data Anal. (this issue), doi:10.1016/j.csda.2009.10.002.

Croux, C., Gallopoulos, E., Van Aelst, S. and Zha, H., 2007. Machine learning and robust data mining. Comp. Statist. Data Anal., 52, 151-154.

Croux, C., Gelper, S. and Mahieu, K., 2010. Robust exponential smoothing of multivariate time series. Comp. Statist. Data Anal. (this issue), doi:10.1016/j.csda.2009.05.003.

Debruyne, M., Hubert, M., Van Horebeek, J., 2010. Detecting influential observations in Kernel PCA. Comp. Statist. Data Anal. (this issue), doi:10.1016/j.csda.2009.08.018.

Daye, Z.J. and Jeng, X.J., 2009. Shrinkage and model selection with correlated variables via weighted fusion. Comp. Statist. Data Anal., 53, 1284-1298.

Dupuis Lozeron, E. and Victoria-Feser, M.P., 2010. Robust estimation of constrained covariance matrices for confirmatory factor analysis. Comp. Statist. Data Anal. (this issue), doi:10.1016/j.csda.2009.08.014.

Emura, T., Lin, C.-W. and Wang, W., 2010. A goodness-of-fit test for Archimedean copula models in the presence of right censoring. Comp. Statist. Data Anal. (this issue), doi:10.1016/j.csda.2010.03.013.

Flores, S., 2010. On the efficient computation of robust regression estimators. Comp. Statist. Data Anal. (this issue), doi:10.1016/j.csda.2010.03.020.

Frahm, G. and Jaekel, U., 2010. A generalization of Tyler's M-estimators to the case of incomplete data. Comp. Statist. Data Anal., 54, 374-393.

García-Escudero, L.A., Gordaliza, A., Mayo-Iscar, A. and San Martín, R., 2010. Robust clusterwise linear regression through trimming. Comp. Statist. Data Anal. (this issue), doi:10.1016/j.csda.2009.07.002.

Gottard, A. and Pacillo, S., 2010. Robust concentration graph model selection. Comp. Statist. Data Anal. (this issue), doi:10.1016/j.csda.2008.11.021.

Haight, T.J., Wang, Y., van der Laan, M.J. and Tager, I.B., 2010. A cross-validation deletion-substitution-addition model selection algorithm: Application to marginal structural models. Comp. Statist. Data Anal. (this issue), doi:10.1016/j.csda.2010.02.002.

Harrington, J. and Salibián-Barrera, M., 2010. Finding approximate solutions to combinatorial problems with very large data sets using BIRCH. Comp. Statist. Data Anal., 54, 655-667.

Hron, K., Templ, M. and Filzmoser P., 2010. Imputation of missing values for compositional data using classical and robust methods. Comp. Statist. Data Anal. (this issue), doi:10.1016/j.csda.2009.11.023.

Hubert, M., Rousseeuw, P. and Verdonck, T., 2009. Robust PCA for skewed data and its outlier map. Comp. Statist. Data Anal., 53, 2264-2274.

Jurečková, J., Picek, J. and Saleh, A.K.Md.E., 2010. Rank tests and regression rank score tests in measurement error models. Comp. Statist. Data Anal. (this issue), doi:10.1016/j.csda.2009.08.020.

Khan, J.A., Van Aelst, S. and Zamar, R.H., 2010. Fast robust estimation of prediction error based on resampling. Comp. Statist. Data Anal. (this issue), doi:10.1016/j.csda.2010.01.031.

Lanius, V. and Gather, U., 2010. Robust online signal extraction from multivariate time series. Comp. Statist. Data Anal., 53, 966-975.

Lee, J.S. and Cox, D.D., 2010. Robust smoothing: Smoothing parameter selection and applications to fluorescence spectroscopy. Comp. Statist. Data Anal. (this issue), doi:10.1016/j.csda.2009.08.001.

Lu, Z.-H., 2009. Covariate selection in mixture models with the censored response variable. Comp. Statist. Data Anal., 53, 2710-2723.

Lutz, R.W., Kalisch, M. and Bühlmann, P., 2008. Robustified L2 boosting. Comp. Statist. Data Anal., 52, 3331-3341.

Lykou, A. and Whittaker, J., 2010. Sparse CCA using a Lasso with positivity constraints. Comp. Statist. Data Anal. (this issue), doi:10.1016/j.csda.2009.08.002.

Maronna, R.A. and Yohai, V.J., 2010. Correcting MM estimates for "fat" data sets. Comp. Statist. Data Anal. (this issue), doi:10.1016/j.csda.2009.09.015.

Maugis, C., Celeux, G. and Martin-Magniette, M.-L., 2009. Variable selection in model-based clustering: A general variable role modeling. Comp. Statist. Data Anal., 53, 3872-3882.

Meintanis, S.G. and Tsionas, E., 2010. Testing for the generalized normal-Laplace distribution with applications. Comp. Statist. Data Anal. (this issue), doi:10.1016/j.csda.2009.05.015.

Menjoge, R.S. and Welsch, R.E., 2010. A diagnostic method for simultaneous feature selection and outlier identification in linear regression. Comp. Statist. Data Anal. (this issue), doi:10.1016/j.csda.2010.02.014.

Min, A., Holzmann, H. and Czado, C., 2010. Model selection strategies for identifying most relevant covariates in homoscedastic linear models. Comp. Statist. Data Anal. (this issue), doi:10.1016/j.csda.2009.09.006.

Nguyen, T.D. and Welsch, R., 2010. Outlier detection and least trimmed squares approximation using semi-definite programming. Comp. Statist. Data Anal. (this issue), doi:10.1016/j.csda.2009.09.037.

Nott, D.J. and Leng, C., 2010. Bayesian projection approaches to variable selection in generalized linear models. Comp. Statist. Data Anal. (this issue), doi:10.1016/j.csda.2010.01.036.

Nunkesser, R. and Morell, O., 2010. An evolutionary algorithm for robust regression. Comp. Statist. Data Anal. (this issue), doi:10.1016/j.csda.2010.04.017.

Ogura, T., 2010. A variable selection method in principal canonical correlation analysis. Comp. Statist. Data Anal., 54, 1117-1123.

Ouysse, R. and Kohn, R., 2010. Bayesian variable selection and model averaging in the arbitrage pricing theory model. Comp. Statist. Data Anal. (this issue), doi:10.1016/j.csda.2009.09.034.

Overstall, A.M. and Forster, J.J., 2010. Default Bayesian model determination methods for generalised linear mixed models. Comp. Statist. Data Anal. (this issue), doi:10.1016/j.csda.2010.03.008.

Powers, S., Gerlach, R. and Stamey, J., 2010. Bayesian variable selection for Poisson regression with underreported responses. Comp. Statist. Data Anal. (this issue), doi:10.1016/j.csda.2010.04.003.

Riani, M. and Atkinson, A.C., 2010. Robust model selection with flexible trimming. Computational Statistics and Data Analysis. Comp. Statist. Data Anal. (this issue), doi:10.1016/j.csda.2010.03.007.

Rozenholc, Y., Mildenberger,T. and Gather, U., 2010. Combining regular and irregular histograms by penalized likelihood. Comp. Statist. Data Anal. (this issue), doi:10.1016/j.csda.2010.04.021.

Rufo, M.J., Martín, J. and Pérez, C.J., 2010. New approaches to compute Bayes factor in finite mixture models. Comp. Statist. Data Anal. (this issue), doi:10.1016/j.csda.2010.05.002.

Salibian-Barrera, M. and Van Aelst, S. 2008. Robust model selection using fast and robust bootstrap. Comp. Statist. Data Anal., 52, 5121-5135.

Schomaker, M., Wan, A.T.K. and Heumann, C., 2010. Frequentist Model Averaging with missing observations. Comp. Statist. Data Anal. (this issue), doi:10.1016/j.csda.2009.07.023

Schyns, M. , Haesbroeck, G. and Critchley, F., 2010. RelaxMCD: Smooth optimisation for the Minimum Covariance Determinant estimator. Comp. Statist. Data Anal., 54, 843-857.

Todorov, V. and Filzmoser, P., 2010. Robust statistic for the one-way MANOVA. Comp. Statist. Data Anal., 54, 37-48.

Unkel, S. and Trendafilov, N.T., 2010. A majorization algorithm for simultaneous parameter estimation in robust exploratory factor analysis. Comp. Statist. Data Anal. (this issue), doi:10.1016/j.csda.2010.02.003.

Wang, H.-B., 2009. Bayesian estimation and variable selection for single index models. Comp. Statist. Data Anal., 53, 2617-2627.

Wang, Y.-G. and Hin, L.-Y., 2010. Modeling strategies in longitudinal data analysis: Covariate, variance function and correlation structure selection. Comp. Statist. Data Anal. (this issue), doi:10.1016/j.csda.2009.11.006.

Wiens, D.P. and Wu, E.K.H., 2010. A comparative study of robust designs for M-estimated regression models. Comp. Statist. Data Anal., 54, 1683-1695.

Stefan Van Aelst, Ghent University, Belgium
E-mail: Stefan.VanAelst@UGent.be

Roy Welsch, Massachusetts Institute of Technology, USA
E-mail: welsch@mit.edu

Ruben H. Zamar, University of British Columbia, Canada
E-mail: ruben@stat.ubc.ca